REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 167–185

NONPARAMETRIC PREDICTIVE INFERENCE FOR REPRODUCIBILITY OF TWO BASIC TESTS BASED ON ORDER STATISTICS

Authors: Frank P.A. Coolen
– Department of Mathematical Sciences, Durham University, Durham, UK
frank.coolen@durham.ac.uk

Hana N. Alqifari
– Department of Mathematics, Qassim University, Buraidah, Saudi Arabia

Received: February 2017 Revised: June 2017 Accepted: July 2017

Abstract:

- Reproducibility of statistical hypothesis tests is an issue of major importance in applied statistics: if the test were repeated, would the same overall conclusion be reached, that is rejection or non-rejection of the null hypothesis? Nonparametric predictive inference (NPI) provides a natural framework for such inferences, as its explicitly predictive nature fits well with the core problem formulation of a repeat of the test in the future. NPI is a frequentist statistics method using relatively few assumptions, made possible by the use of lower and upper probabilities. For inference on reproducibility of statistical tests, NPI provides lower and upper reproducibility probabilities (RP). In this paper, the NPI-RP method is presented for two basic tests using order statistics, namely a test for a specific value for a population quantile and a precedence test for comparison of data from two populations, as typically used for experiments involving lifetime data if one wishes to conclude before all observations are available.

Key-Words:

- *lower and upper probabilities; nonparametric predictive inference; precedence test; quantile test; reproducibility.*

AMS Subject Classification:

- 60A99, 62G99, 62P30.

1. INTRODUCTION

Testing of hypotheses is one of the main tools in statistics and crucial in many applications. While many different tests have been developed for a wide range of scenarios, the aspect of reproducibility of tests has long been neglected: the question addressed is whether or not a test, if it were repeated under the same circumstances, would lead to the same overall conclusion, that is rejection or non-rejection of the null hypothesis. Recently, this topic has started to gain attention, in particular through the publication of a 'handbook on reproducibility' [4] which provides a collection of papers on the issue. Nevertheless, whilst hypothesis testing is mainly seen as a frequentist statistics procedure, the classic frequentist framework is not suited for inference on reproducibility as this is neither an estimation nor a testing problem. The very nature of reproducibility is predictive, namely given the results of one test one wishes to predict the outcome of a possible future test. Coolen and Bin Himd [11] presented nonparametric predictive inference (NPI) for reproducibility of some basic tests, with more attention to this topic in the PhD thesis of Bin Himd [8], these publications also provide a critical discussion of earlier methods for reproducibility presented in the literature.

This paper contributes to development of NPI for reproducibility by considering two tests based on order statistics, namely a one sample quantile test and a two sample precedence test. Central to these inferences are NPI results for future order statistics [12]. This paper provides a concise presentation of NPI for the quantile and basic precedence test, further details, examples and discussion are included in the PhD thesis of Alqifari [1].

This paper is organized as follows. Section 2 provides a brief introduction to NPI, including key results on NPI for future order statistics as used in this paper. Section 3 discusses aspects of reproducibility of statistical tests and explains the NPI perspective on such inferences. Section 4 presents the NPI approach to reproducibility of a basic quantile test. Section 5 considers a precedence test used for comparison of two populations. Some concluding remarks are given in Section 6. All computations in this paper were performed using the statistical software R.

2. NONPARAMETRIC PREDICTIVE INFERENCE

Nonparametric predictive inference (NPI) [5, 10] is a statistical framework which uses few modelling assumptions, with inferences explicitly in terms of future observations. For real-valued random quantities attention has thus far been

mostly restricted to a single future observation, although multiple future observations have been considered for some NPI methods, e.g. in statistical process control [2, 3].

Assume that we have real-valued ordered data $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, with $n \geq 1$. For ease of notation, define $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$, or at other known lower and upper bounds of the range of possible values for these random quantities. The n observations create a partition of the real-line into $n+1$ intervals $I_j = (x_{(j-1)}, x_{(j)})$ for $j = 1, ..., n+1$. We assume throughout this paper that ties do not occur. If we wish to allow ties, also between past and future observations, we could use closed intervals $[x_{(j-1)}, x_{(j)}]$ instead of these open intervals I_j, the difference is rather minimal and to keep presentation easy we have opted not to do this here. We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, ..., m$. We link the data and future observations via Hill's assumption $A_{(n)}$ [17], or, more precisely, via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, ..., m-2$; we will refer to this generically as 'the $A_{(n)}$ assumptions'), which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. The $A_{(n)}$ assumptions imply that all possible orderings of the n data observations and the m future observations are equally likely, where the n data observations are not distinguished among each other and neither are the m future observations. Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, ..., m\}$, then the $A_{(n)}$ assumptions lead to

$$P\left(\bigcap_{j=1}^{n+1} \{S_j = s_j\}\right) = \binom{n+m}{n}^{-1} \tag{2.1}$$

where s_j are non-negative integers with $\sum_{j=1}^{n+1} s_j = m$. Another convenient way to interpret the $A_{(n)}$ assumptions with n data observations and m future observations is to think that n randomly chosen observations out of all $n+m$ real-valued observations are revealed, following which you wish to make inferences about the m unrevealed observations. The $A_{(n)}$ assumptions then imply that one has no information about whether specific values of neighbouring revealed observations make it less or more likely that a future observation falls in between them. For any event involving the m future observations, Equation (2.1) implies that we can count the number of such orderings for which this event holds. Generally in NPI a lower probability for the event of interest is derived by counting all orderings for which this event has to hold, while the corresponding upper probability is derived by counting all orderings for which this event can hold [5, 10].

In NPI, the $A_{(n)}$ assumptions justify the use of resulting inferences directly as predictive probabilities. Using only precise probabilities, such inferences cannot be used for many events of interest, but in NPI we use the fact, in line with De Finetti's Fundamental Theorem of Probability [14], that corresponding optimal bounds can be derived for all events of interest [5]. These bounds are lower and upper probabilities in the theory of imprecise probability [6]. NPI provides

exactly calibrated frequentist inferences [18], and it has strong consistency properties in theory of interval probability [5]. In NPI the n observations are explicitly used through the $A_{(n)}$ assumptions, yet as there is no use of conditioning as in the Bayesian framework, we do not use an explicit notation to indicate this use of the data. The m future observations must be assumed to result from the same sampling method as the n data observations in order to have full exchangeability. NPI is totally based on the $A_{(n)}$ assumptions, which however should be considered with care as they imply e.g. that the specific ordering in which the data appeared is irrelevant, so accepting $A_{(n)}$ implies an exchangeability judgement for the n observations. It is attractive that the appropriateness of this approach can be decided upon after the n observations have become available. NPI is always in line with inferences based on empirical distributions, which is an attractive property when aiming at objectivity [10].

Let $X_{(r)}$, for $r = 1, ..., m$, be the r-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, ..., m$ and $X_{(1)} < X_{(2)} < \cdots < X_{(m)}$. The following probabilities are derived by counting the relevant orderings and use of Equation (2.1). For $j = 1, ..., n+1$ and $r = 1, ..., m$,

$$P\big(X_{(r)} \in I_j\big) = \binom{j+r-2}{j-1}\binom{n-j+1+m-r}{n-j+1}\binom{n+m}{n}^{-1}. \tag{2.2}$$

For this event NPI provides a precise probability, as each of the $\binom{n+m}{n}$ equally likely orderings of n past and m future observations has the r-th ordered future observation in precisely one interval I_j. As Equation (2.2) only specifies the probabilities for the events that $X_{(r)}$ belongs to intervals I_j, it can be considered to provide a partial specification of a probability distribution for $X_{(r)}$, no assumptions are made about the distribution of the probability masses within such intervals I_j.

Analysis of the probability in Equation (2.2) leads to some interesting results, including the logical symmetry $P(X_{(r)} \in I_j) = P(X_{(m+1-r)} \in I_{n+2-j})$. For all r, the probability for $X_{(r)} \in I_j$ is unimodal in j, with the maximum probability assigned to interval I_{j^*} with $\left(\frac{r-1}{m-1}\right)(n+1) \leq j^* \leq \left(\frac{r-1}{m-1}\right)(n+1)+1$. A further interesting property occurs for the special case where the number of future observations is equal to the number of data observations, so $m = n$. In this case, $P(X_{(r)} < x_r) = P(X_{(r)} > x_r) = 0.5$ holds for all $r = 1, ..., m$. This fact can be proven by considering all $\binom{2n}{n}$ equally likely orderings, where clearly in precisely half of these orderings the r-th future observation occurs before the r-th data observation due to the overall exchangeability assumption. The special case $m = n$ plays an important role in this paper as it naturally occurs in analysis of reproducibility of statistical hypothesis tests.

3. REPRODUCIBILITY OF STATISTICAL TESTS

Statistical hypothesis testing is used in many application areas and normally results in either non-rejection of the stated null hypothesis or its rejection in favour of a stated alternative, at a predetermined level of significance. Whilst this procedure is embedded in the successful long-standing tradition of statistics, a related aspect that had received relatively little attention in the literature until recently is the reproducibility of such tests: if the test were repeated, would it lead to the same overall conclusion? Attention to problems with reproducibility, including problems with understanding of concepts by practitioners in application areas, was raised by Goodman [16] and Senn [21]. Methods for addressing reproducibility, proposed in the literature since then, have mainly shown that the classical frequentist framework of statistics may not be immediately suitable for inference on test reproducibility (see [11] for a discussion of such proposed methods). Recently, many aspects of reproducibility, including some attention to statistical methods, have been discussed in a volume dedicated to this topic [4].

The reproducibility probability (RP) for a test is the probability for the event that, if the test is repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is either rejection of the null-hypothesis or not, will be the same. In practice, focus may often be on reproducibility of tests in which the null-hypothesis is rejected, for example because significant effects tend to lead to new treatments in medical applications. However, also if the null-hypothesis is not rejected it is important to have a meaningful assessment of the reproducibility of the test. Note that RP is assessed knowing the outcome of the first, actual experiment, which consists of the actual observations, so not only the value of a sufficient test statistic or even just the conclusion on rejection or non-rejection of the null-hypothesis. This is important as the RP will vary with different experiment outcomes, which is logical and will lead to higher RP if the data supported the original test conclusion more strongly. A sufficient test statistic, if of reduced dimension compared to the full data set, does not provide suitable input for the NPI method, hence the use of the full data set is required for the inferences considered in this paper.

Coolen and Bin Himd [11] introduced NPI for RP, denoted by NPI-RP, by considering some basic nonparametric tests: the sign test, Wilcoxon's signed rank test, and the two sample rank sum test. For these inferences NPI for Bernoulli quantities [9] and for real-valued observations [5] were used. This did not lead to precise valued reproducibility probabilities but to NPI lower and upper reproducibility probabilities, denoted by $\underline{RP}$ and $\overline{RP}$, respectively. For these tests analytic methods were presented to calculate the NPI lower and upper probabilities for test reproducibility. To enable NPI for more complex test scenarios, the NPI-bootstrap method can be used, as introduced and illustrated by Bin Himd [8] for the Kolmogorov–Smirnov test.

This paper presents NPI-RP for two classical tests which are based on order statistics, namely a one sample quantile test (Section 4) and a two sample precedence test (Section 5). For these inferences, NPI for future order statistics [12] is used, as briefly reviewed in Section 2. We assume that the first, actual experiment led to ordered real-valued observations $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. As we consider an imaginary repeat of this experiment, we use NPI for $m = n$ future ordered observations, henceforth denoted by $X^f_{(1)} < X^f_{(2)} < \cdots < X^f_{(n)}$, with the superscript f used to emphasize that we consider future order statistics.

4. QUANTILE TEST

The quantile test is a basic nonparametric test for the value of a population quantile [15]. Let κ_p denote the $100 \times p$-th quantile of an unspecified continuous distribution, for $0 \leq p \leq 1$. On the basis of a sample of observations of independent and identically distributed random quantities X_i, $i = 1, ..., n$, we consider the one-sided test of null-hypothesis H_0: $\kappa_p = \kappa_p^0$ versus alternative H_1: $\kappa_p > \kappa_p^0$, for a specified value κ_p^0. We restrict attention in this paper to NPI for reproducibility of this one-sided quantile test. The corresponding methodology for the two-sided test follows the same steps and is included in the PhD thesis of Alqifari [1], where also some more discussion and examples are given of the tests presented in this paper. Actually, there is an interesting issue about two-sided tests in such scenarios, that requires some further thought. If the original test leads to rejection of the null hypothesis due to a relatively small value of the test statistic, would one consider the test result to be reproduced if a future test leads to rejection due to a relatively large value of the test statistic, so in the other tail of the statistic's distribution under H_0? Technically perhaps this is the case, but on the basis of the combined evidence of the two tests one would probably want to investigate the whole setting further and not regard the second test as confirming the results of the first test. This is left as a topic for consideration.

Under H_0, κ_p^0 is the $100 \times p$-th quantile of the distribution function of the X_i, so $P(X_i \leq \kappa_p^0 \,|\, H_0) = p$. Define the random variable K as the number of X_i in the sample of size n that are less than or equal to κ_p^0, that is

$$K = \sum_{i=1}^{n} \mathbf{1}\{X_i \leq \kappa_p^0\}$$

with $\mathbf{1}\{A\} = 1$ if A is true and $\mathbf{1}\{A\} = 0$ if A is not true. A logical test rule is to reject H_0 if $X_{(r)} > \kappa_p^0$, where $X_{(r)}$ is the r-th ordered observation in the sample (ordered from small to large), for a suitable value of r corresponding to a chosen significance level, so if $K \leq r - 1$. For significance level α, r is the largest integer

such that

$$P\big(X_{(r)} > \kappa_p^0 \,|\, H_0\big) \;=\; \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \;\leq\; \alpha \;.$$

For large sample sizes the Normal distribution approximation to the Binomial distribution can be used in order to determine the appropriate value of r.

For a given data set $x_1, ..., x_n$, the test statistic of the one-sided quantile test as defined above is the number of observations less than or equal to κ_p^0, denoted by

$$k \;=\; \sum_{i=1}^{n} \mathbf{1}\big\{x_i \leq \kappa_p^0\big\} \;.$$

For the value r derived as discussed above, H_0 is rejected if and only if $k \leq r-1$.

Based on such data and the result of the actual hypothesis test, that is whether the null hypothesis is rejected in favour of the alternative hypothesis or not, NPI can be applied to study the reproducibility of the test. First we consider the case where $k \leq r-1$, so the original test leads to rejection of H_0. Reproducibility of this test result is therefore the event that, if the test were repeated, also leading to n observations, then that would also lead to rejection of H_0. Using the notation for future observations introduced in Section 3, this would occur if the r-th ordered observation of the future sample exceeds κ_p^0. The NPI lower and upper reproducibility probabilities for this event, as function of $k \leq r-1$, are

$$\underline{RP}(k) \;=\; \underline{P}\big(X_{(r)}^f > \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_{j-1} > \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big)$$

and

$$\overline{RP}(k) \;=\; \overline{P}\big(X_{(r)}^f > \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_j > \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big) \;,$$

respectively. Note that the dependence of these lower and upper probabilities on the value k is not explicit in the notation used for the terms on the right-hand side, but is due to the number of data x_j that exceed κ_p^0. It is easily shown that $\underline{P}(X_{(r)}^f > \kappa_p^0 \,|\, k) = \overline{P}(X_{(r)}^f > \kappa_p^0 \,|\, k+1)$, which leads to $\underline{RP}(k) = \overline{RP}(k+1)$ for values of k leading to rejection of H_0.

If the original test does not lead to rejection of H_0, so if $k \geq r$, then reproducibility of the test is the event that the null hypothesis would also not get rejected in the future test. The NPI lower and upper reproducibility probabilities for this event, as function of $k \geq r$, are

$$\underline{RP}(k) \;=\; \underline{P}\big(X_{(r)}^f \leq \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_j \leq \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big)$$

and

$$\overline{RP}(k) \;=\; \overline{P}\big(X^f_{(r)} \le \kappa^0_p \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_{j-1} \le \kappa^0_p\big\}\, P\big(X^f_{(r)} \in I_j\big)\;,$$

respectively. It is easily seen that $\underline{RP}(k) = \overline{RP}(k-1)$ for values of k such that $k-1$ leads to H_0 not being rejected. If an actual observation in the original test is exactly equal to the specified value κ^0_p, then the NPI method would actually provide a precise reproducibility probability. We do not consider this further as the test hypotheses must always be specified without consideration of the actual test data, hence this case is extremely unlikely to occur; for some further discussion see [1].

The minimum value that can occur for the NPI lower reproducibility probabilities for this one-sided quantile test, following either rejection or non-rejection of the null hypothesis in the original test, is equal to 0.5. This follows directly from the formulae for the NPI lower reproducibility probabilities given above, together with $P(X_{(r)} < x_r) = P(X_{(r)} > x_r) = 0.5$ as explained in Section 2. The NPI upper reproducibility probabilities can be equal to one. This occurs when all observations in the original test are less than κ^0_p, so $k = n$, in which case the original test let to H_0 not being rejected for all values of r (so for all order statistics considered, hence for any level of significance); this reflects that, with no evidence in the original data in favour of the possibility that the data values can actually exceed κ^0_p, one cannot exclude the possibility that no future observations could exceed this value. Note that the corresponding NPI lower reproducibility probability will be less than one, reflecting that the original data set only provides limited information, this lower probability will increase towards one as function of n. The upper reproducibility probability is also equal to one if all observations in the original test are greater than κ^0_p, so $k = 0$, for which case the reasoning is similar to that above but of course now with H_0 being rejected.

Example 1. Suppose that the original test has sample size $n = 15$ and we are interested in testing the null hypothesis that the third quartile, so the 75% quantile, of the underlying distribution is equal to a specified value $\kappa^0_{0.75}$ against the alternative hypothesis that this third quartile is greater than $\kappa^0_{0.75}$, tested at significance level $\alpha = 0.05$. Using the Binomial distribution for the classical quantile test, this leads to the rule that H_0 is rejected if $x_{(8)} > \kappa^0_{0.75}$ and H_0 is not rejected if $x_{(8)} < \kappa^0_{0.75}$. Note that we do not discuss the case $x_{(8)} = \kappa^0_{0.75}$ which is slightly different as the NPI approach leads to precise probabilities instead of lower and upper probabilities (see [1]), it is also of little practical relevance.

Table 1 presents the NPI lower and upper reproducibility probabilities for all values of k, which is the number of observations in the original test which are less than $\kappa^0_{0.75}$. If $k \le 7$ then the original test leads to H_0 being rejected while it is not rejected for $k \ge 8$. Hence, the NPI lower and upper reproducibility probabilities are for the events $X^f_{(8)} > \kappa^0_{0.75}$ and $X^f_{(8)} < \kappa^0_{0.75}$, respectively. This

table illustrates the logical fact that the worst reproducibility is achieved for k at the threshold values 7 and 8, with increasing RP values when moving away from these values, leading to maximum NPI-RP values for $k = 0$ and $k = 15$. Because for this test the threshold between rejecting and not rejecting H_0 is between $k = 7$ and $k = 8$ out of $n = 15$ observations, the NPI-RP values are symmetric, that is the same for $k = j$ and $k = 15 - j$ for $j = 0, 1, ..., 7$ in Table 1.

Table 1: NPI-RP for third quartile, $n = 15$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	0.9989	1	6	0.6424	0.7689	12	0.9359	0.9749
1	0.9929	0.9989	7	0.5	0.6424	13	0.9749	0.9929
2	0.9749	0.9929	8	0.5	0.6424	14	0.9929	0.9989
3	0.9359	0.9749	9	0.6424	0.7689	15	0.9989	1
4	0.8682	0.9359	10	0.7689	0.8682			
5	0.7689	0.8682	11	0.8682	0.9359			

Table 2 presents NPI-RP values for the quantile test considering the median, so the 50% quantile, again with sample size $n = 15$ and testing the null hypothesis that the median is equal to a specified value $\kappa^0_{0.5}$ against the one-sided hypothesis that it is greater than $\kappa^0_{0.5}$, at level of significance $\alpha = 0.05$. This leads to the test rule that H_0 is rejected if the number k of observations that are smaller than $\kappa^0_{0.5}$ is less than or equal to 3, and H_0 is not rejected if $k \geq 4$. Note that throughout this paper, precise values 0.5 and 1 are presented without additional decimals, so the values 1.0000 are actually less than 1 but rounded upwards. Of course, these NPI-RP values are not symmetric, and reproducibility becomes very likely for initial test results with a substantial number of observations less than $\kappa^0_{0.5}$. But rejection of H_0, which occurs for $k \leq 3$ and is often of main practical relevance, has relatively low NPI-RP values.

Table 2: NPI-RP for median, $n = 15$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	0.9502	1	6	0.7865	0.8775	12	0.9986	0.9997
1	0.8352	0.9502	7	0.8775	0.9359	13	0.9997	0.9999
2	0.6743	0.8352	8	0.9359	0.9698	14	0.9999	1.0000
3	0.5	0.6743	9	0.9698	0.9873	15	1.0000	1
4	0.5	0.6592	10	0.9873	0.9954			
5	0.6592	0.7865	11	0.9954	0.9986			

Tables 3 and 4 present the NPI-RP results for the same one-sided quantile test on the third quartile for $n = 30$, at significance levels $\alpha = 0.05$ and $\alpha =$

0.01, respectively. Using the Normal distribution approximation, the test rule for $\alpha = 0.05$ is to reject H_0 that this third quartile is equal to $\kappa^0_{0.75}$ in favour of the alternative hypothesis that it is greater than $\kappa^0_{0.75}$ if $k \leq 18$ and not to reject it if $k \geq 19$, where k is again the number of observations less than $\kappa^0_{0.75}$. For $\alpha = 0.01$, H_0 is rejected if $k \leq 16$ and not rejected if $k \geq 17$. The change in level of significance α leads obviously to change of the rejection threshold, with H_0 being rejected for a smaller range of values k in case of smaller value of α. Comparison of Table 3 with Table 1 shows that the larger sample size tends to lead to slightly less imprecision, that is the difference between corresponding upper and lower probabilities, this is e.g. shown by considering the upper probabilities $\overline{RP}(k)$ for the values of k next to the rejection thresholds, so corresponding to $\underline{RP}(k) = 0.5$.

Table 3: NPI-RP for third quartile, $n = 30$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	1.0000	1	11	0.9651	0.9811	22	0.7941	0.8666
1	1.0000	1.0000	12	0.9398	0.9651	23	0.8666	0.9210
2	1.0000	1.0000	13	0.9023	0.9398	24	0.9210	0.9580
3	1.0000	1.0000	14	0.8503	0.9023	25	0.9580	0.9805
4	0.9999	1.0000	15	0.7826	0.8503	26	0.9805	0.9923
5	0.9998	0.9999	16	0.6995	0.7826	27	0.9923	0.9976
6	0.9993	0.9998	17	0.6038	0.6995	28	0.9976	0.9995
7	0.9981	0.9993	18	0.5	0.6038	29	0.9995	0.9999
8	0.9956	0.9981	19	0.5	0.6054	30	0.9999	1
9	0.9905	0.9956	20	0.6054	0.7056			
10	0.9811	0.9905	21	0.7056	0.7941			

Table 4: NPI-RP for third quartile, $n = 30$ and $\alpha = 0.01$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	1.0000	1	11	0.9023	0.9406	22	0.9101	0.9483
1	1.0000	1.0000	12	0.8493	0.9023	23	0.9483	0.9731
2	1.0000	1.0000	13	0.7805	0.8493	24	0.9731	0.9875
3	0.9999	1.0000	14	0.6971	0.7805	25	0.9875	0.9949
4	0.9995	0.9999	15	0.6019	0.6971	26	0.9949	0.9983
5	0.9986	0.9995	16	0.5	0.6019	27	0.9983	0.9995
6	0.9964	0.9986	17	0.5	0.6026	28	0.9995	0.9999
7	0.9916	0.9964	18	0.6026	0.6995	29	0.9999	1.0000
8	0.9824	0.9916	19	0.6995	0.7852	30	1.0000	1
9	0.9664	0.9824	20	0.7852	0.8559			
10	0.9406	0.9664	21	0.8559	0.9101			

5. PRECEDENCE TEST

As a second example of NPI for reproducibility of a statistical test based on order statistics we consider a basic nonparametric precedence test. Such a test, first proposed by Nelson [19], is typically used for comparison of two groups of lifetime data, where one wishes to reach a conclusion before all units on test have failed. The test is based on the order of the observed failure times for the two groups, and typically leads to, possibly many, right-censored observations at the time when the test is ended. Balakrishnan and Ng [7] present a detailed introduction and overview of precedence testing, including more sophisticated tests than the basic one considered in this paper. NPI for precedence testing was presented by Coolen-Schrijner *et al.* [13], without consideration of reproducibility. It should be emphasized that we consider here the NPI approach for reproducibility of a classical precedence test, so not of the NPI approach to precedence testing [13].

We consider the classical scenario with two independent samples. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n_x)}$ be the ordered real-valued observations in a sample of size n_x drawn randomly from a continuously distributed population, which we refer to as the X population, with a probability distribution depending on location parameter λ_x. Similarly, let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n_y)}$ be the ordered real-valued observations in a sample of size n_y drawn randomly from another continuously distributed population (the Y population) with a probability distribution which is identical to that of the X population except for its location parameter λ_y. The hypothesis test for the locations of these two populations considered here is $H_0 \colon \lambda_x = \lambda_y$ versus $H_1 \colon \lambda_x < \lambda_y$, which is to be interpreted such that, under H_1, observations from the Y population tend to be larger than observations from the X population.

The precedence test considered in this paper, for this specific hypothesis test scenario, is as follows. Given n_x and n_y, one specifies the value of r, such that the test is ended at, or before, the r-th observation of the Y population. For specific level of significance α, one determines the value k (which therefore is a function of α and of r) such that H_0 is rejected if and only if $X_{(k)} < Y_{(r)}$. The critical value for k is the smallest integer which satisfies

$$P\big(X_k < Y_r | H_0\big) \; = \; \binom{n_x + n_y}{n_x}^{-1} \sum_{j=0}^{r-1} \binom{j+k-1}{j} \binom{n_y - j + n_x - k}{n_y - j} \; \leq \; \alpha \, .$$

Note that the test is typically ended at the time $T = \min(X_{(k)}, Y_{(r)})$, with the conclusion that H_0 is rejected in favour of the one-sided alternative hypothesis H_1 specified above if $T = X_{(k)}$ and H_0 is not rejected if $T = Y_{(r)}$. It is of interest to emphasize this censoring; continuing with the original test would make no difference at all to the test conclusion, but further observations would make a difference for the NPI reproducibility results, as will be discussed later.

The NPI approach for reproducibility of this two-sample precedence test considers again the same test scenario applied to future order statistics, and derives the lower and upper probabilities for the event that the same overall test conclusion will be derived, given the data from the original test. This involves the earlier described NPI approach for inference on the r-th future order statistic $Y^f_{(r)}$ out of n_y future observations based on the data from the Y population, and similarly for the k-th future order statistics $X^f_{(k)}$ out of the n_x future observations based on the data from the X population, where the values of r and k are the same as used for the original test (as we assume also the same significance level for the future test). Note, however, that there is a complication: for full specification of the NPI probabilities for these future order statistics, we require the full data from the original test to be available. But, as mentioned, the data resulting from the original precedence test typically has right-censored observations for at least one, but most likely both populations, and these are all just known to exceed the time T at which the original test had ended.

Before we proceed, we discuss this situation in more detail as it is important for the general idea of studying reproducibility of tests. We should emphasize that we have not come across this issue before in the literature, but it seems to be important and more details are provided by Alqifari [1]. There are two perspectives on the study of reproducibility of such precedence tests. First, one can study the test outcome assuming that actually complete data were available, so all n_x and n_y observations of the X and Y populations, respectively, in the original test are assumed to be available. Secondly, one can consider inference for the realistic scenario with the actual data from the original test, so including right-censored observations at time T. The first scenario is the most straightforward for the development of NPI-RP, and we start with this scenario. Then we explain how this first scenario, without additional assumptions, leads to NPI-RP for the second scenario.

The starting point for NPI-RP for the precedence test is to apply NPI for n_x future observations, based on the n_x original test observations from the X population, which are assumed to be fully available, and similarly for n_y future observations based on the n_y observations from the Y population. Using the results presented in Section 2, with notation adapted to indicate the specific populations, the following NPI lower and upper reproducibility probabilities are derived. First, if H_0 is rejected in the original test, so $x_{(k)} < y_{(r)}$, then

$$\underline{RP} = \underline{P}\big(X^f_{(k)} < Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x)} < y_{(j_y-1)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,,$$

$$\overline{RP} = \overline{P}\big(X^f_{(k)} < Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x-1)} < y_{(j_y)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,.$$

If H_0 is not rejected in the original test, so $x_{(k)} > y_{(r)}$, then

$$\underline{RP} = \underline{P}\big(X^f_{(k)} > Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x-1)} < y_{(j_y)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,,$$

$$\overline{RP} = \overline{P}\big(X^f_{(k)} > Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x)} < y_{(j_y-1)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,.$$

The following general results for this NPI lower and upper reproducibility probabilities are easily derived [1]. Both in case of rejecting and not rejecting H_0, the maximum possible value of the NPI upper reproducibility probability is 1. If H_0 was rejected this occurs if $x_{(n_x)} < y_{(1)}$, while if H_0 was not rejected this occurs if $x_{(1)} > y_{(n_y)}$, so both cases lead to maximum reproducibility if the original test data were entirely separated in the sense that either all observations from X population occurred before all observations from the Y population, or the other way around.

In both cases of rejecting or not rejecting H_0 in the original test, the minimum value of the NPI lower reproducibility probability is 0.25. If H_0 was rejected, this occurs if $y_{(r-1)} < x_{(1)}$ and $x_{(k)} < y_{(r)}$ and $y_{(n_y)} < x_{(k+1)}$. If H_0 was not rejected, this occurs if $x_{(k-1)} < y_{(1)}$ and $y_{(r)} < x_{(k)}$ and $x_{(n_x)} < y_{(r+1)}$. Both these smallest possible values for $\underline{RP}$ result from data orderings that, whilst leading to a test conclusion, are least supportive for it, together with the fact that $P(X^f_{(k)} < x_{(k)}) = P(X^f_{(k)} > x_{(k)}) = 0.5$ (and similar for $Y^f_{(r)}$) as discussed in Section 2.

The effect of local changes to the combined ordering of the data of the two populations in the original test is important. Suppose that, for given data for the X and Y populations for the original test, observations $y_{(u)}$ and $x_{(v)}$ are such that $y_{(u)} < x_{(v)}$ and in the combined ordering of all $n_x + n_y$ data they are consecutive. Now suppose that we change these observations, and denote them by $\tilde{y}_{(u)}$ and $\tilde{x}_{(v)}$, respectively, such that they keep their order in the data from their own population but between them change their order, so $\tilde{x}_{(v)} < \tilde{y}_{(u)}$. Then this local change to the combined ordering of the data leads to increase of both the NPI lower and upper probabilities for the event $X_{(k)} < Y_{(r)}$, that is

$$\underline{P}\big(X_{(k)} < Y_{(r)} \,|\, y_{(u)} < x_{(v)}\big) \;<\; \underline{P}\big(X_{(k)} < Y_{(r)} \,|\, \tilde{x}_{(v)} < \tilde{y}_{(u)}\big)\,,$$

$$\overline{P}\big(X_{(k)} < Y_{(r)} \,|\, y_{(u)} < x_{(v)}\big) \;<\; \overline{P}\big(X_{(k)} < Y_{(r)} \,|\, \tilde{x}_{(v)} < \tilde{y}_{(u)}\big)\,.$$

This implies that the NPI-RP inferences for the precedence test depend monotonically on the combined ordering of the original test data, which is an important property to derive such inference for actual tests including right-censored observations, as discussed after the next example.

Example 2. Nelson [20] presents data consisting of six groups of times (in minutes) to breakdown of an insulating fluid subjected to different levels of voltage. To illustrate NPI-RP for the basic precedence test as discussed above, we assume that sample 3 provides data from the X population and sample 6 from the Y population, these times are presented in Table 5. Both samples are of size 10, and we assume that the precedence testing scenario discussed in this section is followed, so we assume that the population distributions may only differ in location parameters, with $H_0\colon \lambda_x = \lambda_y$ tested versus $H_1\colon \lambda_x < \lambda_y$. We assume that $r = 6$, so the test is set up to end at the observation of the sixth failure time for the Y population. We discuss both significance levels $\alpha = 0.05$ and $\alpha = 0.1$. The missing values in Table 5 are only known to exceed 3.83.

Table 5: Times to insulating fluid breakdown.

X sample	0.94	0.64	0.82	0.93	1.08	1.99	2.06	2.15	2.57	*
Y sample	1.34	1.49	1.56	2.10	2.12	3.83	*	*	*	*

For significance level $\alpha = 0.05$, the critical value is $k = 10$, while for $\alpha = 0.1$ this is $k = 9$. Therefore, the provided data will lead, in this precedence test, to rejection of H_0 at 10% level of significance but not to rejection of H_0 at 5% level of significance. For both scenarios, the NPI lower and upper reproducibility probabilities are presented in Table 6, for all of the possible orderings of the right-censored observations. Note that in total 15 observations are available, with 1 value of the X sample and 4 values of the Y sample only known to exceed 3.83.

Table 6: NPI-RP for precedence test on insulating fluid breakdown data.

rank of x_{10}	$\alpha = 0.05$		$\alpha = 0.1$	
	$\underline{RP}$	$\overline{RP}$	$\underline{RP}$	$\overline{RP}$
16	0.3871	0.7814	0.3885	0.7079
17	0.4746	0.8209	0.3490	0.6665
18	0.5496	0.8484	0.3215	0.6309
19	0.6019	0.8627	0.3072	0.6062
20	0.6290	0.8669	0.3029	0.5934

In this table, we give the rank, from the combined ordering of all 20 observations, of the right-censored observation $x_{(10)}$, for example when this is 17 it implies that $y_{(7)} < x_{(10)} < y_{(8)}$. Table 6 presents both the results for $\alpha = 0.05$, in which case H_0 was not rejected in the original test, hence reproducibility is achieved if H_0 is also not rejected in the future test, and the results for $\alpha = 0.05$, in which case

H_0 was rejected so reproducibility also implies rejection of H_0 in the future test. Note that for $\alpha = 0.1$ we still assume that $y_{(6)} = 3.83$ was actually observed, even though the test could have been concluded at time $x_{(9)} = 2.57$ because $x_{(9)} < y_{(6)}$ was conclusive for the test in this case. Table 6 shows that the NPI-RP values are increasing in the combined rank of $x_{(10)}$ for $\alpha = 0.05$ and decreasing for $\alpha = 0.1$, which illustrates the monotonicity of these inferences with regard to changes in ranks of the data as discussed above, as increasing combined rank of $x_{(10)}$ provides more evidence in support of H_0, hence in favour of reproducing the original test result for $\alpha = 0.05$ but against doing so for $\alpha = 0.1$. We notice that the actual rank that $x_{(10)}$ would have among the 20 combined observations has substantial influence on the NPI-RP values. In this example, the imprecision $\overline{RP} - \underline{RP}$ is large. This is due to the relatively small data sets and the fact that two groups of data are compared, with imprecision for the predictive inferences for both groups through the $A_{(n)}$ assumptions for each group.

Thus far, we have studied reproducibility of the basic precedence test from the perspective of having the complete data available, in Example 2 this was illustrated by considering all possible orderings for the right-censored data in the two samples. However, a more realistic perspective is to only use the actual test outcome, without any assumptions on the ordering of the right-censored observations. Using lower and upper probabilities, this can be easily achieved by defining $\underline{RP}$ as the minimum of all NPI lower probabilities for reproducibility over all possible orderings for the right-censored observations, and similarly by defining $\overline{RP}$ as the maximum of all NPI upper probabilities for reproducibility over all possible orderings for the right-censored observations. Hence, in Example 2, this leads to $\underline{RP} = 0.3871$ and $\overline{RP} = 0.8669$ for $\alpha = 0.05$, and $\underline{RP} = 0.3029$ and $\overline{RP} = 0.7079$ for $\alpha = 0.1$. Of course, this leads to increased imprecision compared to every possible specific ordering of the right-censored observations, but it is convenient as no further assumptions about those right-censored observations are required. Furthermore, to derive the NPI-RP values for this perspective one does not need to calculate the corresponding values for each possible combined ordering of right-censored observations, due to the above discussed monotonicity of these inferences. Hence, we always know for which specific ordering of right-censored observations these NPI-RP values are obtained, that is either with all right-censored observations from the X sample occurring before all right-censored observations from the Y sample, or the other way around, depending on the actual outcome of the original test. This perspective is illustrated further in Example 3.

Example 3. We consider again NPI-RP for the precedence test as presented in this section, so with one-sided alternative hypothesis H_1: $\lambda_x < \lambda_y$. Suppose that $n_x = 10$ units of the X population and $n_y = 8$ units of the Y population are put on a life test, where one wants at most two Y units to actually fail, so the value $r = 2$ is chosen. Testing at significance level $\alpha = 0.05$, the critical value is $k = 7$, so H_0 is rejected if $x_{(7)} < y_{(2)}$ while H_0 is not rejected if $y_{(2)} < x_{(7)}$.

Note that, with the test ending at time $\min(x_{(7)}, y_{(2)})$, there are least 3 right-censored X observations and at least 6 right-censored Y observations; this leads to large imprecision in the NPI-RP values. Table 7 presents the NPI lower and upper reproducibility probabilities for this test, for all possible data in the original test, which are indicated through the rankings of all observations until the test is ended, in the combined ranking of the X and Y samples. As indicated, the columns to the left relate to the cases where H_0 is not rejected while the columns to the right relate to the cases where H_0 is rejected. All these NPI-RP values are calculated using the monotonicity with regard to the combined ranks of the right-censored observations, as explained above. These results illustrate the earlier discussed maximum value 1 for $\overline{RP}$ and minimum value 0.25 for $\underline{RP}$. It is particularly noticeable that the NPI lower reproducibility probabilities for this test tend to be small, which is not really surprising due to the large number of right-censored observations resulting from the choice $r = 2$.

Table 7: NPI-RP for precedence test with $n_x = 10$, $n_y = 8$, $r = 2$, $k = 7$ and $\alpha = 0.05$.

H_0 not rejected				H_0 rejected			
X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$	X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$
—	1,2	0.4992	1	1–7	—	0.3833	1
1	2,3	0.4951	0.9988	1–6,8	7	0.3367	0.8833
2	1,3	0.4970	0.9992	1–5,7,8	6	0.2993	0.8425
1,2	3,4	0.4826	0.9924	1–4,6–8	5	0.2739	0.8098
1,3	2,4	0.4884	0.9946	1–3,5–8	4	0.2593	0.7875
2,3	1,4	0.4903	0.9951	1,2,4–8	3	0.2526	0.7748
1–3	4,5	0.4553	0.9733	1,3–8	2	0.2504	0.7690
1–4	5,6	0.4075	0.9314	2–8	1	0.25	0.7670
1–5	6,7	0.3375	0.8582				
1–6	7,8	0.25	0.7509				
2–7	1,8	0.3663	0.8375				

6. CONCLUDING REMARKS

The NPI approach to reproducibility of tests provides many research challenges. It can be developed for many statistical tests, while for some data types (e.g. multivariate data) first NPI requires to be developed further. The test scenarios studied for particular tests may require careful attention, as illustrated by the different perspectives discussed for the precedence test in Section 5. As mentioned, the precedence test scenario discussed in this paper is very basic. Balakrishnan and Ng [7] present a detailed introduction and overview of precedence testing, including more sophisticated tests than the basic one considered

in this paper. In practice, it is important for such tests, and also in general, to also consider the power of the test; thus far this has not yet been considered in the NPI approach for reproducibility of testing. With further development of this approach, we are aiming at guidance on selection of test methods which, for specified level of significance, have good power and good reproducibility properties. This may often require more test data than needed following traditional guidance, but the assurance of good reproducibility is important for many applications and may lead to savings in the longer run by reducing processes, such as development of new medication, to continue on the basis of false test results which may later turn out not to be reproduced in repeated tests under similar circumstances. Further details, examples and discussion of the tests presented in this paper are given in the PhD thesis of Alqifari [1].

ACKNOWLEDGMENTS

This work was carried out while Hana Alqifari was studying for PhD at the Department of Mathematical Sciences, Durham University, supported by the Ministry of Higher Education in Saudi Arabia and Qassim University. The authors gratefully acknowledge detailed comments by three anonymous reviewers which led to improved presentation of the paper. The authors also thank Professor Sat Gupta for the kind invitation to present this work at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics at The University of North Carolina at Greensboro (October 2016), and the subsequent invitation to submit this paper for the conference special issue of this journal.

REFERENCES

[1] Alqifari, H.N. (2017). *Nonparametric Predictive Inference for Future Order Statistics*, PhD Thesis, Durham University (available from `www.npi-statistics.com`).

[2] Arts, G.R.J. and Coolen, F.P.A. (2008). Two nonparametric predictive control charts, *Journal of Statistical Theory and Practice*, **2**, 499–512.

[3] Arts, G.R.J.; Coolen, F.P.A. and van der Laan, P. (2004). Nonparametric predictive inference in statistical process control, *Quality Technology and Quantitative Management*, **1**, 201–216.

[4] Atmanspacher, H. and Maasen, S. (Eds.) (2016). *Reproducibility: Principles, Problems, Practices and Prospects*, Wiley, Hoboken, New Jersey.

[5] Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability, *Journal of Statistical Planning and Inference*, **124**, 251–272.

[6] Augustin, T.; Coolen, F.P.A.; de Cooman, G. and Troffaes, M.C.M. (Eds.) (2014). *Introduction to Imprecise Probabilities*, Wiley, Chichester.

[7] Balakrishnan, N. and Ng, H.K.T. (2006). *Precedence-Type Tests and Applications*, Wiley, Hoboken, New Jersey.

[8] Bin Himd, S. (2014). *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility*, PhD Thesis, Durham University (available from `www.npi-statistics.com`).

[9] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem, *Statistics and Probability Letters*, **36**, 349–357.

[10] Coolen, F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism, *Journal of Logic, Language and Information*, **15**, 21–47.

[11] Coolen, F.P.A. and Bin Himd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests, *Journal of Statistical Theory and Practice*, **8**, 591–618.

[12] Coolen, F.P.A.; Coolen-Maturi, T. and Alqifari, H.N. (2018). Nonparametric predictive inference for future order statistics, *Communications in Statistics — Theory and Methods*, **47**, 2527–2548.

[13] Coolen-Schrijner, P.; Maturi, T.A. and Coolen, F.P.A. (2009). Nonparametric predictive precedence testing for two groups, *Journal of Statistical Theory and Practice*, **3**, 273–287.

[14] De Finetti, B. (1974). *Theory of Probability* (2 volumes), Wiley, London.

[15] Gibbons, J.D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference* (5th ed.), Chapman & Hall, Boca Raton, FL.

[16] Goodman, S.N. (1992). A comment on replication, p-values and evidence, *Statistics in Medicine*, **11**, 875–879.

[17] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population, *Journal of the American Statistical Association*, **63**, 677–691.

[18] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions, *Biometrika*, **92**, 529–542.

[19] Nelson, L.S. (1963). Tables for a precedence life test, *Technometrics*, **124**, 491–499.

[20] Nelson, W.B. (1982). *Applied Life Data Analysis*, Wiley, Hoboken, New Jersey.

[21] Senn, S. (2002). Comment on 'A comment on replication, p-values and evidence', by S.N. Goodman (Letter to the editor), *Statistics in Medicine*, **21**, 2437–2444. With author's reply, pp. 245–247.